Towards the Automatic Acquisition of Lexical Data

H.Trost, E.Buchberger
Department of Medical Cybernetics and Artificial Intelligence
University of Vienna, Austria

Abstract
Creating a knowledge base has always been a bottleneck in the implementation of AI systems. This is also true for Natural Language Understanding (NLU) systems, particularly for data-driven ones. While a perfect system for automatic acquisition of all sorts of knowledge is still far from being realized, partial solutions are possible. This holds especially for lexical data. Nevertheless, the task is not trivial, in particular when dealing with languages rich in inflectional forms like German. Our system is to be used by persons with no specific linguistic knowledge, thus linguistic expertise has been put into the system to ascertain correct classification of words. Classification is done by means of a small rule based system with lexical knowledge and language-specific heuristics. The key idea is the identification of three sorts of knowledge which are processed distinctly and the optimal use of knowledge already contained in the existing lexicon.

1. Introduction

In this paper we introduce a system for the semi-automatic enlargement of a morphological lexicon. It forms part of VIE-LANG, a German language dialogue system (Buchberger et al. 1982). VIE-LANG serves not only as an object but as a meta system as well: its knowledge base is to be enlarged, and its facilities are used to support that process: the parser serves to analyze the input to the acquisition system, the generator is used to provide examples.

In contrast to English the morphological analysis of German words is no trivial task, due to two causes:
- First, there is a rich inflectional system, consisting of about 60 different endings (where most endings have various different interpretations), some prefixes ('ge-' for PPP, 'zu' for some infinitive forms), 'umlautung', and irregular forms,
- second, lemmatization has to be complemented by interpretation, because the functional structure of a sentence is not bound to constituent order but to the case system (expressed by inflection) instead.

To build up a lexicon, one needs a classification of German words on the basis of their graphemic realization. There exist several works on this subject, e.g. Bergmann (1982), Knopik (1984), Schott (1978), Schulze and Heinze (1982), Willee (1979). For VIE-LANG we developed our own classification scheme (Trost and Dorffner 1986), based on Kunze and Ruediger (1968).

For all those schemes it takes an expert to classify new words correctly. Our acquisition system contains linguistic expertise in the form of different types of rules, which allows for semi-automatic acquisition of lexical knowledge in an interaction with a user who need not have specific linguistic knowledge.

Whereas different approaches for knowledge acquisition for NLU systems have been proposed (e.g. Ballard (1984), Haas and Hendrix (1982)), we concentrate on the acquisition of lexical data for the German language by using specific properties of this domain.

2. The Morphologic Classification Scheme

Our classification scheme is based primarily on the sets of suffixes that can be attached to certain stems. Every different set constitutes a morphological class, and every lexicon entry falls exactly into one of these classes. Altogether there are about 70 different ones. For each class two lists are stored: One containing the set of suffixes belonging to the class, and another one containing the syntactic interpretation for each of these suffixes. Superimposed on this scheme is information about 'umlaut' and prefix 'ge'. They occur only in a few positions, depending on the word category. Every possible combination is represented by a certain numerical value stored along with each lexicon entry.

We distinguish between morphologic and syntactic information, the latter being a feature of the lexeme itself (and not expressed by inflection). Morphologic information consists of the following features:

KL : The morphologic class as above
UM : Information about 'umlaut'
PF : Information about formation of PPP (verbs only)
FM : Information about other forms (suppletion)

The syntactic information is stored in the feature SY. It consists of the following data:
- word category (verb, noun, pronoun, etc.)
- gender of nouns
- subcategory (auxiliary, modal, proper name, etc.)
- case (for prepositions)
- auxiliary for present and past perfect ('haben' or 'sein')
- separable verbadjuncts

This information is coded into a number, the first digit representing the word category, the other ones depending on it (e.g. gender only for nouns).

As an example let's look at the entries for the verb 'geben' (to give). Three forms are to be considered, 'geb' is the stem for present tense and PPP, 'gib' for 2nd and 3rd person sg present tense indicative, and 'gab' for past tense. The corresponding dictionary entries have the following form:

```
GEB: Key:  LXM#889              GIB: Key:  LXM#718
     KL:   22                        KL:   26
     UM:   0                         UM:   0
     PF:   1                         PF:   0
     FM:   8
     SY:   500                  GAB: Key:  LXM#754
     FORR: (LXM#718 LXM#754)         KL:   23
                                     UM:   3
                                     PF:   0
```

The two lists for morphologic class 22 are given below:

```
END22 : (E EN END EST ET T)
INT22:  (E (111 121 123) EN (3 6 114 124) END (4)
         EST (122)       ET (125)         T (115 52))
```

The suffix list gives the possible endings of the words in class 22, the interpretation list gives the code of all forms expressed by any one of these endings.

3. Knowledge Base

The acquisition system is rule based. Its knowledge base comprises three types of rules:

- Rules representing inflectional paradigms. These rules describe the basic types of conjugation and declination in German.
- Morphonological rules. The basic inflectional endings are split up into a much larger set by various morphonological rules which alter the endings and stems to make pronunciation easier.
- Heuristic rules. While the former two rule types are derived from the German grammar proper, these rules are like plausible guesses. They guide the system to make choices like which category a word belongs to according to knowledge about forms (i.e. all verbs end with -en), actual frequency of classes, etc.

These rules are organized in distinct packages. Only rules in active packages are considered. Rules may activate and deactivate rule packages.

4. Overall Architecture

According to their different nature, the three mentioned types of rules are processed differently. Knowledge about inflectional types serves to partition the words into disjunct classes. Once the inflectional type has been determined, there are relatively clear guidelines as to the inflection of the word. The inflectional type actually is a subclassification of the word type.

One of the crucial points is determining the word type. The system first tries to make use of its basic vocabulary. It checks whether a new word is composed of words already in the lexicon or of an existing word stem together with a derivational ending. There is a rule in German morphology stating that in compound words the morphological class is determined by the last word. On a similar line reasoning about derivational endings is performed, as those may determine word type as well as inflection. As a next heuristic morphological clues are taken into consideration. There exist a number of them, but ambiguities may arise. If this is the case, a third strategy is applied: the system asks the user to type in a short utterance containing the new word. The utterance is analysed by the parser of VIE-LANG rendering information about the word type by means of the phrase type it appears in. In applying this method, the system relies on a simple but important presupposition: the user usually enters an utterance containing the word in a proper linguistic context facilitating determination of its type. We do not argue that the user will always utter the minimal projection, but that he will not violate phrase borders with his utterance. The knowledge about phrase types as well as the basic vocabulary permits unambiguous determination of the word type in most cases, especially as the most irregular forms that are very limited in number (words of the closed word classes: pronouns, articles, auxiliary and modal verbs, etc.) have already been included in the basic lexicon.

Once the word type has been determined, the rule package associated with it is activated. Let's suppose the new word is a verb. Then, the verb-package is triggered. Here in turn we find packages for strong and weak inflection. The large number of subclasses is implied by morphonological reasons, whereby the small number of general paradigms is multiplied. Morphonologic rules have exact matching conditions, therefore classification in this part is automated to a large extent. The only problem is deciding for weak or strong inflection first. As exact rules do not exist, heuristics are applied which are based mainly on word frequency.

An important feature is the dynamic interaction register: the hypotheses evoked by the heuristic rules require to be confirmed by the user. The system knows which word forms will form sufficient evidence for a certain hypothesis. It will generate these forms and ask the user for confirmation. The forms however depend on the hypotheses. Thus, the user is only asked a minimum of questions. The forms to be asked for are kept in a dynamic interaction register which is updated with every hypothesis and every answer from the user.

5. An Example Session

In this chapter we show how a new entry is actually created. The user starts the interaction by entering a new word, e.g. 'abgeben' (to leave). The first thing the system has to do is to decide about the word category. To find out if it is a compound word it will try to split off words first from the beginning then from the end.

This will result in recognizing 'ab' as a separable verbadjunct. Of course the 'ab' could be part of a totally different stem like 'Abend' (evening) or 'aber' (but). So the system looks for facts supporting the verb hypothesis. Verbs are usually typed in in infinitive form and this implies the ending '-en' (in a few cases also '-n'). Of course this '-en' could also be part of a stem like 'Magen' (stomach) or 'wegen' (because), but the combination of both verb adjunct 'ab' and ending '-en' on a word belonging to a different category is highly unplausible. So 'abgeben' is split into ab/geb/en.

As a next step the lexicon is looked up for 'geb'. If it is found the rest is easy. All the information from 'geb' is simply duplicated; the only additional information to be stored is about the separable 'ab'. This way the new entry may be created without any other help by the user.

To continue with our example we will assume that 'geb' is not already contained in the lexicon. That means the system has to figure out a hypothesis concerning the conjugation type of 'abgeben' (either weak or strong). Since weak verbs make up the vast majority of German verbs, this hypothesis is tried first.

	FORM	CLASS	FM	UM	PF	SY
present tense	abgeb	44	0	0	1	502

Weak conjugation is regular, all forms are built from one stem. To confirm weak conjugation it suffices to show the user the 1st person sg past tense. Before doing so all morphonological rules connected to weak conjugation are tried. None applies, so user interaction can start. 1st person sg of past tense in the weak paradigm is 'gebte ab'. To make sure the user knows which form is intended, some context has to be provided. This leads to the phrase 'gestern gebte ich ab' (I leaved yesterday) specifying tense and person. The user recognizes 'gebte' as incorrect and rejects that phrase. This makes the system discard the hypothesis weak and try strong instead.

Strong conjugation is more complicated than weak. There may be a maximum of four different stems for

present tense, present tense 2nd and 3rd person sg, past tense and PPP. All these possibilities have either to be resolved automatically or asked explicitly from the user. First the system continues to determine the past tense forms. There are three different types of vowel changes in the case of 'e'-stems (e-a-e, e-o-o, e-a-o). They are sorted by frequency, because no other criterion is available. Again all morphonological rules applicable to strong verbs are tried. In our case none applies, so the user is asked again for verification with 'gestern gab ich ab' (I left yesterday).

	FORM	CLASS	FM	UM	PF	SY
present tense	abgeb	30				
pres.t.2nd p.sg						
past tense	abgab	23				
past participle						

This time the user confirms, so the system can go on. There are two possibilities for the PPP, and again the more frequent one is tried, and accepted by the user.

There is still another irregularity concerning 2nd and 3rd person sg present tense. In most of the cases the stem vowel 'e' becomes 'i'. After verification of this fact the morphological class is finally determined. The system creates three lexical entries 'abgeb', 'abgib' and 'abgab' for present and PPP, 2nd and 3rd person sg present tense and past tense respectively.

Now all of the features have to be filled in. PF of 'abgeb' is set to 1, since the verbadjunct 'ab' implies the use of the prefix 'ge-' for the PPP. UM is set to 8 for 'abgab', indicating 'umlautung' for the subjunctive mode in the past tense. FM of the primary entry 'abgeb' is set to 8 as a result of the combination of classes. Then SY is set to 502 (5 = verb, 0 = present perfect with 'haben', 2 = separable verbadjunct of length 2).

	FORM	CLASS	FM	UM	PF	SY
present tense	abgeb	22	8	0	1	502
pres.t.2nd p.sg	abgib	26	-	0	0	-
past tense	abgab	23	-	8	0	-

Next all indicative forms of present and past tense and the PPP are printed and the user is asked for confirmation. This step could actually be skipped but it is another safety measure against faulty entries.

In our specific example there is a final step to be done: Since 'geb' was not found in the lexicon, it has to be included, too, for two reasons. First the analysis algorithm otherwise could not handle all those cases where the particle is actually split off in the text, second there may be more compound verbs with 'geb', and their incorporation into the lexicon can then be handled fully automatic. Since the verb stem of a compound verb with separable verbadjunct can always appear as a verb in its own right, this poses no problem. The situation is slightly more difficult with other particles where this is not granted. In those cases the new entry must be marked as internal, so that it does not affect analysis or synthesis.

Creation of the new entries is simple anyway. All forms are duplicated, 'abgeb', 'abgib' and 'abgab' are changed to 'geb', 'gib', 'gab' respectively and SY is set to 500 instead of 502.

6. Conclusion

We have presented a system which automates acquisition of lexical data for a natural language understanding system to a large extent. Knowledge acquisition takes place in graceful interaction with a human who is not supposed to have specific linguistic knowledge. The system relies on the existing natural language system VIE-LANG containing among other sources of knowledge a lexicon with a basic vocabulary such that acquisition does not start from scratch but can be seen as an iterative process. The acquisition system is based on a small rule based system in which three different sorts of knowledge - inflectional, morphonological and heuristic - are distinguished and processed differently. As for derivational endings as well as compound words the system heavily relies on existing lexicon entries to form its hypotheses.

The described system forms part of an integrated system for the acquisition of different sorts of knowledge for natural language understanding. An outline of the overall system is to be found in Trost and Buchberger (1985). The final goal will be a system which augments its knowledge automatically in every interaction with the user in a practical and comfortable way.

Acknowledgments

Part of this work was sponsored by the Austrian 'Fonds zur Foerderung der wissenschaftlichen Forschung', grant no.5468.

REFERENCES:

Ballard B.W.: The Syntax and Semantics of User-Defined Modifiers in a Transportable Natural Language Processor, in Proceedings of the 10th International Conference on Computational Linguistics, Stanford Univ., California; 1984.

Bergmann H.: Lemmatisierung in HAM-ANS, HAM Memo ANS-10, Universitaet Hamburg; 1982.

Buchberger E., Steinacker I., Trappl R., Trost H., Leinfellner E.: VIE-LANG - A German Language Understanding System, in Trappl R.(ed.), Cybernetics and Systems Research, North-Holland, Amsterdam; 1982.

Haas N., Hendrix G.G.: Learning by Being Told: Acquiring Knowledge for Information Management, in R.S.Michalski et al.(eds.), Machine Learning: An Artificial Intelligence Approach, Tioga, Calif.; 1982

Knopik T.: MORPHY - Die morphologische Komponente zu einem Generierungssystem für das Deutsche, Diplomarbeit, Inst.f.Informatik, Univ.Stuttgart; 1984.

Kunze J., Ruediger B.: Algorithmische Synthese der Flexionsformen des Deutschen, Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 21,245-303; 1968.

Schott G.: Automatische Deflexion deutscher Woerter unter Verwendung eines Minimalwoerterbuchs, Sprache und Datenverarbeitung 1, 62-77; 1978.

Schulze W., Heinze G.: Die Morphosyntaktische Komponente in der Wortdatenbank des Deutschen, Sprache und Datenverarbeitung 1-2,34-42; 1982.

Trost H., Buchberger E.: Knowledge Acquisition in the System VIE-LANG, in H.Trost, J.Retti (eds.), Österreichische Artificial-Intelligence-Tagung 1985, Springer, Berlin; 1985.

Trost H., Dorffner G.: A System for Morphological Analysis and Synthesis of German Texts, in D.Hainline (ed.): New Developments in Computer Assisted Language Learning, Crooms Helm Ltd., London; in print.

Willee G.: LEMMA - Ein Programmsystem zur automatischen Lemmatisierung deutscher Wortformen, Sprache und Datenverarbeitung 1-2,45-60; 1979.